December 2007

Classification Algorithm Sensitivity to Training Data with Non Representative Attribute Noise

Michael Mannino
University of Colorado at Denver

Yanjuan Yang
University of Colorado at Denver and Health Sciences Center

Young Ryu
University of Texas at Dallas

Classification Algorithm Sensitivity to Training Data with Non Representative Attribute Noise

Michael Mannino[1], Yanjuan Yang[1], and Young Ryu[2]

Abstract

We present an empirical comparison of major classification algorithms when training data contains attribute noise levels not representative of field data. Although conventional wisdom indicates that training data should contain noise representative of field data, it can be difficult to ensure representative noise levels. To study classification algorithm sensitivity, we develop an innovative experimental design using noise situation (under or over representation of training noise), algorithm, noise level, and training set size as factors. We consider situations of uniform attribute noise levels on all attributes, variable noise levels, and noise levels assigned by attribute importance. Our results contradict conventional wisdom indicating that investments to achieve representative noise levels may not be worthwhile. In general, over representative training noise should be avoided while under representative training noise is less of a concern. However, the interactions among algorithm, noise level, and training set size indicate that these general results may not apply to particular practice situations.

Keywords: Area under the Receiver Operating Curve, attribute noise, classification algorithm

1. Introduction

Classification algorithms, like other inductive methods, can be sensitive to data quality. In particular, attribute or input noise can have a significant impact on the performance of a classification algorithm. Attribute value errors in a training set can cause a classification algorithm to form a rule with an incorrect state for an input, while errors in cases to be classified can cause the wrong rule to be used. Attribute noise includes errors from incorrectly measuring an input, wrongly reporting the state of an input, relying on stale values, and using imprecise measurement devices. Error rates in large data sets can be larger than 5% unless careful measures are taken to reduce errors (Orr 1998, Redman 1998). Redman (1996) reported error rates for credit records as high as 30% with some recent anecdotal evidence by Pierce and Ackerman (2005) to confirm this high error rate.

The study involves evaluation of asymmetric attribute noise on classification algorithm performance. Asymmetric means that noise levels in training data are significantly different than noise levels when a classifier is deployed in the field. The conventional wisdom on classifier design is to replicate field noise in training data. This research considers situations when noise levels are different. One situation involves training data obtained from experts rather than from historical data. When training data is obtained from historical data, it is difficult to discern whether noise levels in training data are similar to noise encountered in classifier usage. In addition, disruptions in the environment can change the noise levels encountered in the field.

To study asymmetric attribute noise effects, we develop an innovative experimental design with algorithm, noise level, and training set size as factors and relative performance change as the performance measure. We consider under representative training noise (low training noise and high test noise) and over representative training noise (high training noise and low test noise). For noise generation, we study uniform input noise levels on all attributes, variable noise levels, and noise levels assigned by attribute importance. We use two multiple factor research models with repeated measures to test individual factors and factor interactions. To provide a level of external validity, we conduct experiments with four large data sets having diversity of data types, number of attributes, prevalence, and classification difficulty.

Our results indicate that over representative training noise should be avoided while under representative training noise is less of a concern. Cleaning field data improves performance except when a data set is difficult to classify. However, interactions among algorithm, noise level, and training set size indicate that these general results

[1] The Business School, University of Colorado and Health Science Center, Denver, CO 80217, USA,

Michael.Mannino@cudenver.edu and Yanjuan.Yang@cudenver.edu

[2] School of Management, University of Texas at Dallas, Richardson, Texas 75083-0688, YoungRyu@utdallas.edu

may not apply to particular practice situations. In comparisons of relative sensitivity of five prominent classification algorithms to asymmetric noise, the most common result was no significant relative sensitivity. However, interactions among training set size and noise level indicates that differences may occur in practice.

This study has important implications for understanding classification algorithm performance under asymmetric noise. Many business decision environments involve input noise especially involving data provided by outside parties. For example, errors in credit reports, census data, and court records are common. In noisy environments, robustness of algorithm performance can be more important than performance under laboratory conditions. This work is the first systematic study to document differences in classification algorithm robustness under varying levels of asymmetric attribute noise and training set sizes.

This paper is organized as follows. Section 2 presents the experimental design to study the impacts of asymmetric noise. Section 3 analyzes the experimental results and discusses issues related to the results. Section 4 concludes the study.

2. Research Methodology

This section describes the research questions and methodology employed to test the research questions. We describe the hypotheses, framework, experiment design, experiment control procedure, and data sets.

2.1 Research Questions and Framework

The conventional wisdom about attribute noise as established in studies by Quinlan (1986a,b) is that training data should contain noise representative of field data. He demonstrated that the classification accuracy of ID3 was worse for a noise-free training set if the level of field noise is high (45% or greater). A more recent study by Zhu and Wu (2004) indicates that training data should be clean regardless of the noise in field data. Our goal is to extend these studies with a focus on different levels of noise in training and field deployment. As described in the following points, we use a range of classification algorithms, training set sizes, and refined performance measures to extend the results in these previous studies.

- Training data with non representative noise levels: We want to characterize the relationship between noise level and performance degradation. An improved understanding of performance degradation may provide guidance about investment decisions for acquisition of training data. If over representation of noise is harmful, an organization may want to remove excessive noise if field data is relatively clean. If under representation is not harmful, an organization should not expend resources to obtain a training set with noise representative of field data.
- Cleaning field data: Organizations may also consider investments to improve the quality of field data. We want to characterize the relationship between noise level and performance improvements from cleaning of field data. Cleaning field data is more expensive than training data so organizations may want to see significant performance improvements before improving data quality.
- Tolerance of asymmetric noise by popular classification algorithms: Because ensemble methods (Dietterich 2000) have more tolerance for classification noise, we expect that ensemble classifiers will be more tolerant of asymmetric attribute noise than other classifiers. Beyond this expectation, the study will evaluate the sensitivity of popular classification algorithms to different levels of asymmetric attribute noise.
- Interaction of training set size and asymmetric attribute noise: Learning curves and training set size have been carefully studied because of the expense of collecting training data. Interaction of training set size with asymmetric noise is important for justifying investment decisions in training data. We expect to see more sensitivity to asymmetric attribute noise on small training sets than medium and large training sets.
- Impact of noise variation (uniform versus variable) and attribute importance: Previous theoretical studies have demonstrated more harm due to variable noise on important attributes (Laird 1988 and Goldman and Stone 1995). We expect that variable asymmetric attribute noise will lead to more performance degradation than uniform asymmetric noise. Asymmetric attribute noise directed towards important attributes should have more effect than asymmetric attribute noise that is randomly directed.

To study these research questions, we use a framework involving comparisons between different levels of training and testing noise as depicted in Table 2. Test noise levels indicate actual noise encountered in classifier deployment in the field. Our major interest is in situations of under or over representation of noise in training data. For intra-algorithm comparisons, we use absolute performance differences where AUC_{ij} denotes area under the receiver operating curve under the specified levels of training (*i*) and test noise (*j*). For over representative training noise, we use AUC_{HL} - AUC_{LL} (high training noise and low test noise). For under representative training noise, we use $AUC_{LH} – AUC_{HH}$ (low training noise and high test noise). For comparisons across classification algorithms, we use relative performance differences to focus on noise impact, not other differences among classification algorithms. We use $\frac{AUC_{HL} - AUC_{LL}}{AUC_{LL}}$ for over representative training noise levels and $\frac{AUC_{LH} - AUC_{HH}}{AUC_{HH}}$ for under representative training noise levels. Classification algorithms are evaluated for zero and non-zero levels of low noise with the noise level difference (high – low) constant in both cases.

Table 2: Asymmetric Noise Situations

	Test Noise	
Training Noise	*Low*	*High*
Low	AUC_{LL}	AUC_{LH}
High	AUC_{HL}	AUC_{HH}

We are also interested in situations in which cleaning can be done in field data. For intra algorithm comparisons, we use $AUC_{HL} – AUC_{HH}$ for cleaning field data. If cleaning field data improves performance, additional cleaning of training data is covered by the case of over represented training noise.

Noise level difference, training set size, and classification algorithm are used as factors. Because establishing an easily understood functional relationship is difficult, we use discrete noise level differences (low: 0 to 0.05, medium: 0.10 to 0.15, and high: 0.20 to 0.25). We randomly choose the noise level difference from a uniform distribution between the specified end points. Because noise levels are not under control of the data mining professional, we randomly vary the levels rather than setting fixed levels of noise. Two training set sizes are used (low: 200 and high: 1,000). Since training set size is often controllable by the data mining professional, constant values are used. We use five popular classification algorithms having different approaches about concept representation and search: decision tree induction, logistic regression, nearest neighbor, support vector machine, and a meta classifier using bagging and boosting.

Noise levels are assigned to attributes as uniform noise (same noise level on all attributes), variable noise (different noise levels on attributes), and importance sampled noise (noise level selected by attribute importance). For each observation, a noise difference is randomly selected from a uniform distribution between the specified ranges. For uniform noise, zero and non-zero levels of low noise are considered. With zero-level low noise, the high noise level is the noise difference. For non zero-level low noise, the high level is chosen by randomly selecting a value between the noise difference and 25% larger than the noise difference. The low noise level is the high noise level minus the noise difference. For the variable noise variation, each attribute is randomly assigned a noise level from a uniform distribution between the noise difference and the high noise level used in the uniform noise case. The low noise level is the difference between the randomly assigned high noise level and the noise difference. For importance noise variation, the noise levels used in the variable case are sorted in ascending order. Noise levels are assigned according to attribute importance (largest noise level to the most important attribute) or reverse attribute importance (smallest noise level to the most important attribute). Attribute importance is determined by the Ranker search method with information gain as the evaluation function available in Weka (Witten and Frank 2005).

2.2 Experiment Design and Procedures

Our experiment design emphasizes internal validity about conclusions on individual data sets. We use a repeated measures design to investigate noise impacts on individual algorithms (intra-algorithm experiments) and sensitivity to noise among algorithms (inter-algorithm experiments). For each experiment, we repeat the randomly determined factor levels (noise level and training set size). In addition, we control the composition of cases in observations consisting of a pair of a training set and a test set. Given the number of factors and the need to isolate the impact of noise on training and test data, an emphasis on internal validity seems appropriate. We could not achieve a high level of internal validity if we had used the individual data set as an observation. We provide some insights about external validity (across data sets) by repeating the experiments on different data sets.

The primary set of experiments applies to intra-algorithm performance using noise situation, noise level difference, and training set size as factors as shown in Table 3. An observation involves a combination of a training set and a test set chosen by sampling without replacement from a classified data set. A randomly selected noise difference level is applied to the training and/or test sets as indicated by a noise situation (TR_L-TS_L, TR_L-TS_H, TR_H-TS_L, TR_H-TS_H) representing combinations of low (L) and high (H) noise on training (TR) and test (TS) data. Thus, each cell in Table 3 contains 240 observations for the specified training set size and noise level applied to four noise situations. The test set size is two times the training set size. Since we use large data sets, data availability is not a problem for the training and test sets. After generating noise, a classifier is generated using the specified classification algorithm and training set.

An observation is the performance of the specified classifier on the test set. To control for variance, identical training/test sets are used for each noise situation and noise level and identical noise levels are used for each training set size. A separate experiment is conducted for each classification algorithm.

Table 3: Sample Sizes for Intra-Algorithm Experiments

	Noise Level Difference		
Training Set Size **Noise Situation**	*Low*	*Medium*	*High*
200			
TR_L-TS_L	60	60	60
TR_L-TS_H	60	60	60
TR_H-TS_L	60	60	60
TR_H-TS_H	60	60	60
1,000			
TR_L-TS_L	60	60	60
TR_L-TS_H	60	60	60
TR_H-TS_L	60	60	60
TR_H-TS_H	60	60	60

For inter-algorithm comparisons, we use a slightly different design with relative performance difference as the dependent variable. We perform pairwise comparisons among algorithms using a mix of training set sizes (200 and 1,000) and noise level differences (low, medium, and high). We use 60 paired observations for a specified combination of noise level difference and training set size applied to five classification algorithms. Thus, an experiment involves a total of 360 (6 × 60) observations for each classification algorithm. A paired observation involves the same training and testing set for each classification algorithm. A separate experiment is conducted for each noise situation (over-representative and cleaning field data) and data set. Only variable noise variation is used in the inter-algorithm experiments.

To execute the experiments, control software was developed in Microsoft Visual Studio. The control software randomly perturbs training and test data, builds classifiers using training data and selected classification algorithms, classifies test data using the classifiers, and calculates classifier performance. The experiment control program uses classification algorithms available in the Weka software for machine learning (Witten and Frank 2005) and data sets stored as Oracle 10g tables. We selected five popular algorithms available in Weka: J4.8 (Quinlan 1993), AdaBoostM1 (Freund and Schapire 1996), SMO (support vector machine classifier (Keerthi et al. 2001)), IBk (*k* nearest neighbor classifier (Aha and Kibler1991)), and Logistic (Ridge logistic regression (Cessie and Houwelingen 1992)). Separate experiments are executed for the large data sets described in Table 4. The data sets provide a mix of class distributions, data types, and classification difficulty (see Table 5 for average AUC scores with clean data).

Table 4: Summary of Data Sets

Data Set	*Source*	*Characteristics*
Adult	UCI Repository (Hettich et al.1998)	14 attributes, mixed data types, 45,222 cases, about 25-75 class split (income level)
DGP1	Generated by DGP/2 program from UCI repository	10,000 cases, 20 numeric attributes, 60-40 class split, 3 peaks per attribute
Bankruptcy	Bankruptcy data from S&P Compustat North American database	12 numeric attributes, 12212 cases, about 97-3 class split (bankruptcy status)
Thoracic	United Network for Organ Sharing	19 tri state attributes, 13,326 cases, about 67-33 class split (thoracic transplant survival)

Table 5: Average AUC Results on Clean Data for Each Data Set and Algorithm

Dataset\Algorithm	*AdaBoost*	*J4.8*	*IBK*	*Logistic*	*SMO*
Adult	0.868	0.769	0.817	0.832	0.857
Bank	0.786	0.528	0.594	0.747	0.626
DGP	0.548	0.549	0.591	0.530	0.525
Thoracic	0.709	0.653	0.692	0.722	0.713

3. Analysis of Results

This section presents the experiment results and discusses insights from the analysis. The significance of the intra and inter-algorithm models are evaluated followed implications of the study on investment decisions in training data.

3.1 Evaluation of Models

The intra-algorithm experimental results[3] show a complex pattern for under and over representative training noise but a clearer pattern for cleaning field data as shown in Table 6. For the two easier data sets (Adult and Thoracic), over representative training noise ($AUC_{HL} – AUC_{LL}$) is usually significant. High training set size influences significance for SMO (Adult data set) and Logistic (Adult and Thoracic data sets) indicating that small training sets may not capture enough patterns for these algorithms. On the Adult data set, J4.8 is significant only at low noise level differences possibly due to the lower performance on clean data for J4.8 than the other algorithms (Table 5). For the two more difficult data sets (Bank and DGP), over representative training noise is usually not significant. J4.8 is significant for both data sets when training set size and noise level difference are both high, while IBk is significant for the artificial DGP data. The Bank data set is difficult due to its high skew with some algorithms able to cope with the skew. The DGP data is difficult for all algorithms as shown in Table 5.

For three data sets (Adult, Bank, and DGP), under representative training noise ($AUC_{LH} – AUC_{HH}$) is usually not significant. Table 6 shows some exceptions for large training sets and high noise levels especially for IBk. For the Thoracic data set, under representative training noise is usually significant with the exception of J4.8 and SMO. J4.8 is the lowest performing algorithm on the Thoracic data set.

[3] We used a traditional a level (0.05) for post hoc comparison tests. The family error rate (maximum probability that at least one comparison test has a Type I error) is 0.265 using the Bonferroni correction.

Table 6: Summary of Intra-Algorithm Test Results

Dataset | Alg		*Over (training)*	*Under (training)*	*Cleaning (field)*
Adult	J4.8	Significant when noise level is low	Not significant	Significant
	AdaBoost	Significant	Not significant	Significant
	SMO	Significant when TS_H-NL_L[4]	Not significant	Significant
	IBk	Significant	Significant for uniform and variable noise	Significant
	Logistic	Significant when training set size is high.	Significant when training set size is high.	Significant
Bank	J4.8	Significant when TS_H-NL_L	Significant when TS_H-NL_L	Significant when training set size is high
	AdaBoost	Not significant	Not significant	Significant when training set size is high
	SMO	Not significant	Not significant	Not significant
	IBk	Not significant	Not significant	Significant when training set size is high
	Logistic	Not significant	Significant when TS_L-NL_H	Significant when training set size is high
DGP	J4.8	Significant when TS_H-NL_H	Significant when TS_H-NL_H	Significant when training set size is high
	AdaBoost	Not significant	Not significant	Significant
	SMO	Not significant	Not significant	Not significant
	IBk	Significant	Significant when training set size is high	Significant except for TS_L-NL_L and TS_L-NL_M
	Logistic	Not significant	Not significant	Not significant
Thoracic	J4.8	Significant	Not significant	Significant
	AdaBoost	Significant	significant	Significant
	SMO	Significant	Significant when TS_H-NL_H for uniform and importance noise.	Significant
	IBk	Significant	Significant	Significant
	Logistic	Significant when training set size is high for uniform, variable noise	Significant	Significant

The results for cleaning field data partially confirm the results in (Zhu and Wu 2004). Cleaning field data always significantly improves performance for the easier data sets (Adult and Thoracic). For the difficult data sets (Bank and DGP), cleaning field data has mixed results. Cleaning field data is only significant for large training sets for the Bank data set for all algorithms except SMO. The DGP results are highly mixed with SMO and Logistic not significant for any factors and J4.8 only significant with large training sets, and IBk significant for small training set sizes with low and moderate noise level differences.

For the inter-algorithm comparison, we considered two noise situations (over-representative training noise and cleaned field data) with variable noise. For over-representative training noise, a positive performance difference

[4] TS_H-NL_H means high training set size and high noise level difference.

(algorithm 1 – algorithm 2) means that algorithm 1 has less relative sensitivity than algorithm 2. Likewise, a negative performance difference means that algorithm 1 has more relative sensitivity than algorithm 2. Since the level of noise can be difficult to control in a training set, less sensitivity is desired. For cleaned field data, a positive performance difference (algorithm 1 – algorithm 2) means that algorithm 1 has more relative sensitivity than algorithm 2. These results provide insights into the desirability of cleaning field data because cleaning field data for a more sensitive algorithm has more impact than cleaning field data for a less sensitive algorithm.

The inter-algorithm results are mixed as summarized in Table 7. The conclusions in Table 7 are based on significance counts for each data set as shown in Table 8. A Count+ value indicates the number of times a statistical test was significant comparing the algorithm to any other algorithm with a mean difference less sensitive. A Count0 value indicates the number of times a statistical test was not significant comparing the algorithm to any other algorithm. For example in 24 statistical tests with over representative training noise, SMO is less sensitive in 16 tests but not significantly different in 8 tests. Most comparisons for the Adult data set are not significant as the Count0 column counts are largest in all but two rows. Across data sets, IBk is more sensitive on three data sets (Adult, Thoracic, and DGP), while Logistic is more sensitive on the DGP and Thoracic data sets for over representative training noise. For cleaned field data, Logistic regression is most sensitive on three data sets but less sensitive on one data set (DGP).

Table 7: Summary of Inter-Algorithm Test Results

Data Set	*Over representative training*	*Cleaned field data*
Adult	SMO is less sensitive than others. AdaBoost and IBk are more sensitive to noise than others.	Logistic is most sensitive. IBk is somewhat less sensitive than others.
Bank	Most differences are not significant.J4.8 is less sensitive than Logistic when training set size is high and noise level difference is medium.	Algorithms' differences are significant when training set size is high. IBk and Logistic are more sensitive than others. SMO is less sensitive to noise.
DGP	Logistic is less sensitive than others. IBk is most sensitive to noise.	IBk is most sensitive to noise. SMO and Logistic are less sensitive to noise than others.
Thoracic	Logistic is less sensitive than others. IBk is most sensitive to noise.	Logistic is more sensitive to noise. IBk is less sensitive to noise.

Table 8: Significance Counts for the Adult Data Set

Noise Situation	*Algorithm*	*Count +*	*Count –*	*Count 0*
Over	J4.8	0	5	19
	AdaBoost	1	10	13
	SMO	16	0	8
	IBk	2	10	12
	Logistic	9	3	12
Cleaning	J4.8	4	6	14
	AdaBoost	2	8	14
	SMO	3	9	12
	IBk	2	10	12
	Logistic	22	0	2

In many data mining studies, algorithm performance is evaluated using graphs instead of statistical tests. Performance graphs can provide different conclusions than statistical tests as depicted in Figure 1 on the Adult data set. Figure 1 shows stacked bar graphs in which the height of a bar is the sum of the median relative AUC performance differences. J4.8 appears most sensitive in Figure 1 although the statistical tests indicate that the relative performance difference between J4.8 and other algorithms is usually not significant. Likewise, the statistical conclusions for the cleaned field data are not apparent in the bar graphs of Figure 2.

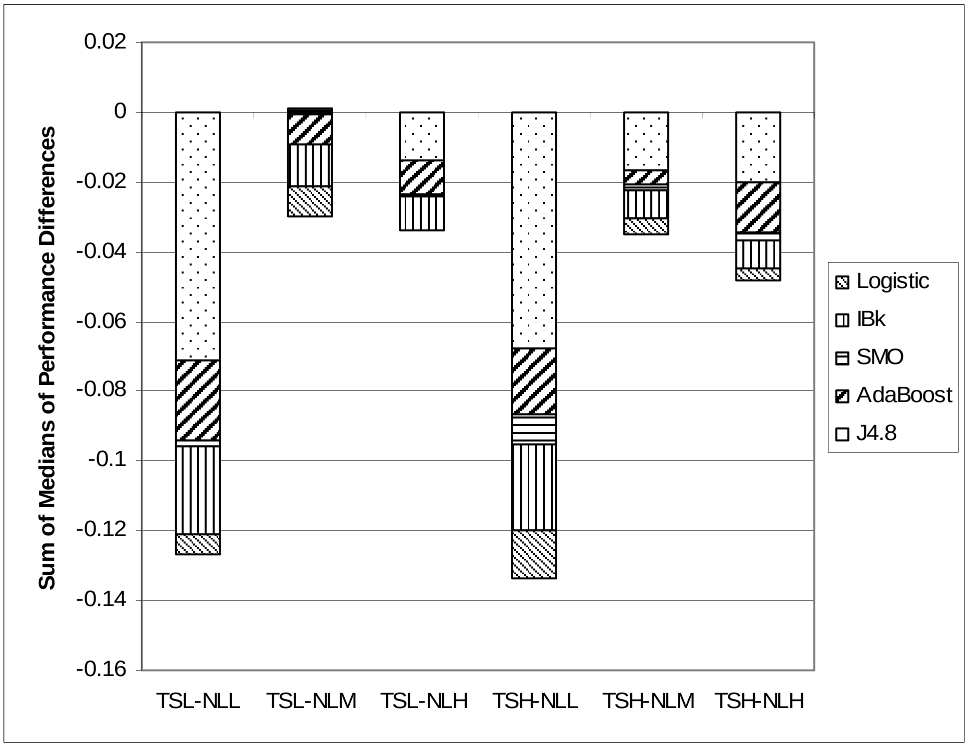

Figure 1: Performance Differences for Over Representative Noise of the Adult Data Set

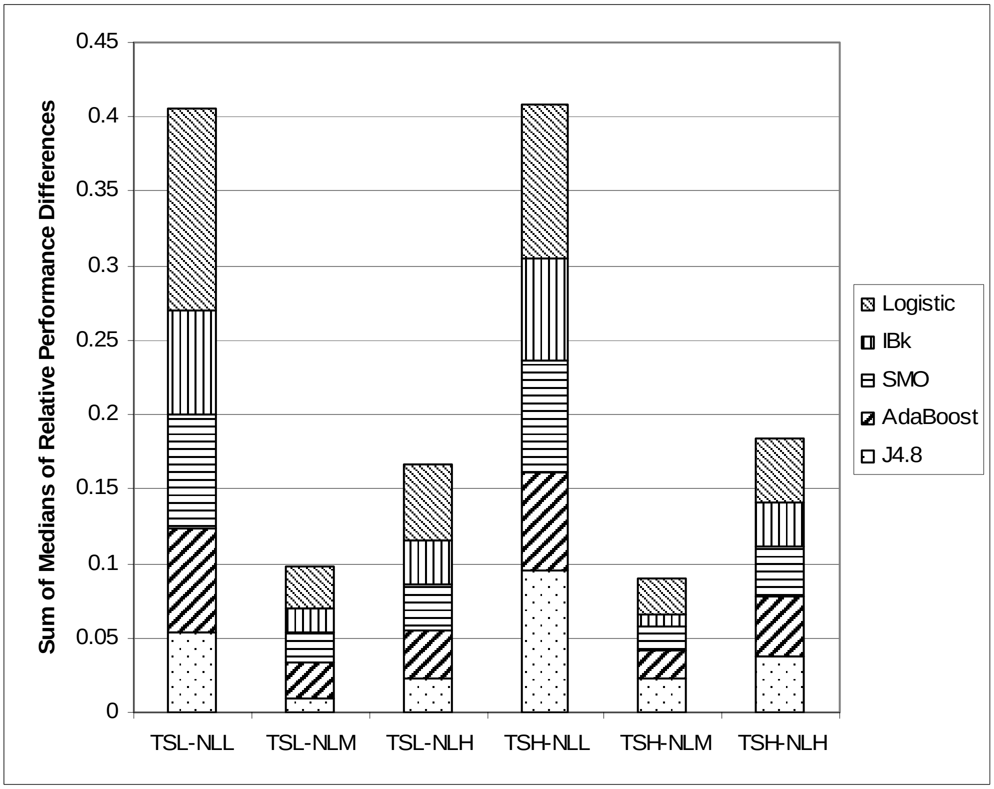

Figure 2: Performance Differences for Cleaned Field Noise of the Adult Data Set

The box charts in Figures 3 and 4 indicate the dispersion of the performance measures in the intra and inter-algorithms experiments. The performance on the Bankruptcy data set shows most dispersion on both parts of the distribution. The wide dispersion may be due to the interaction of noise and high class skew. The box charts also show anomalous situations in which over-representative training noise can lead to improved performance and cleaning field data can lead to worse performance. These anomalous situations are most pronounced for the Bankruptcy data set but exist to lesser degrees for the other data sets.

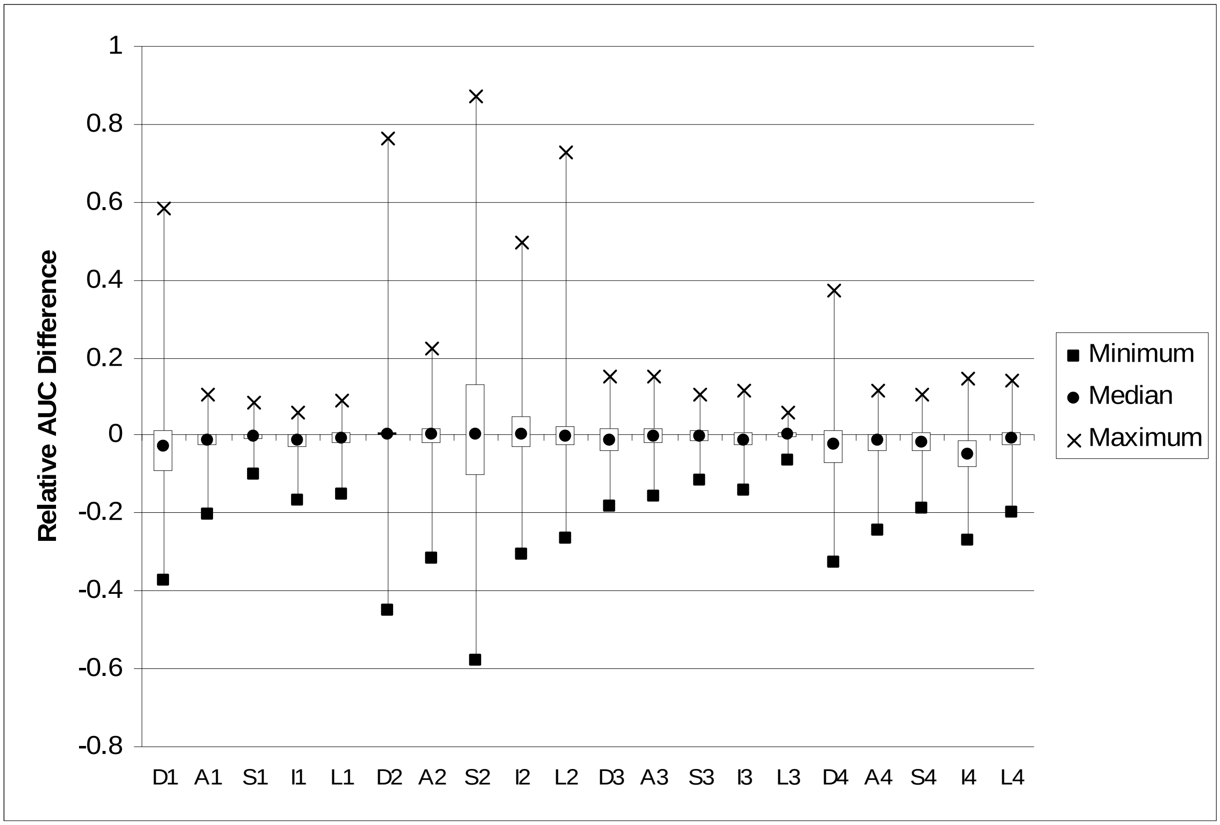

Figure 3: Box Charts for the Over Representative Training Noise Situation[5]

[5] D-J4.8; A-Adaboost; S-SMO; N-IBk; L-Logistic regression; 1-Adult; 2-Bankruptcy; 3-DGP; 4-Thoracic;

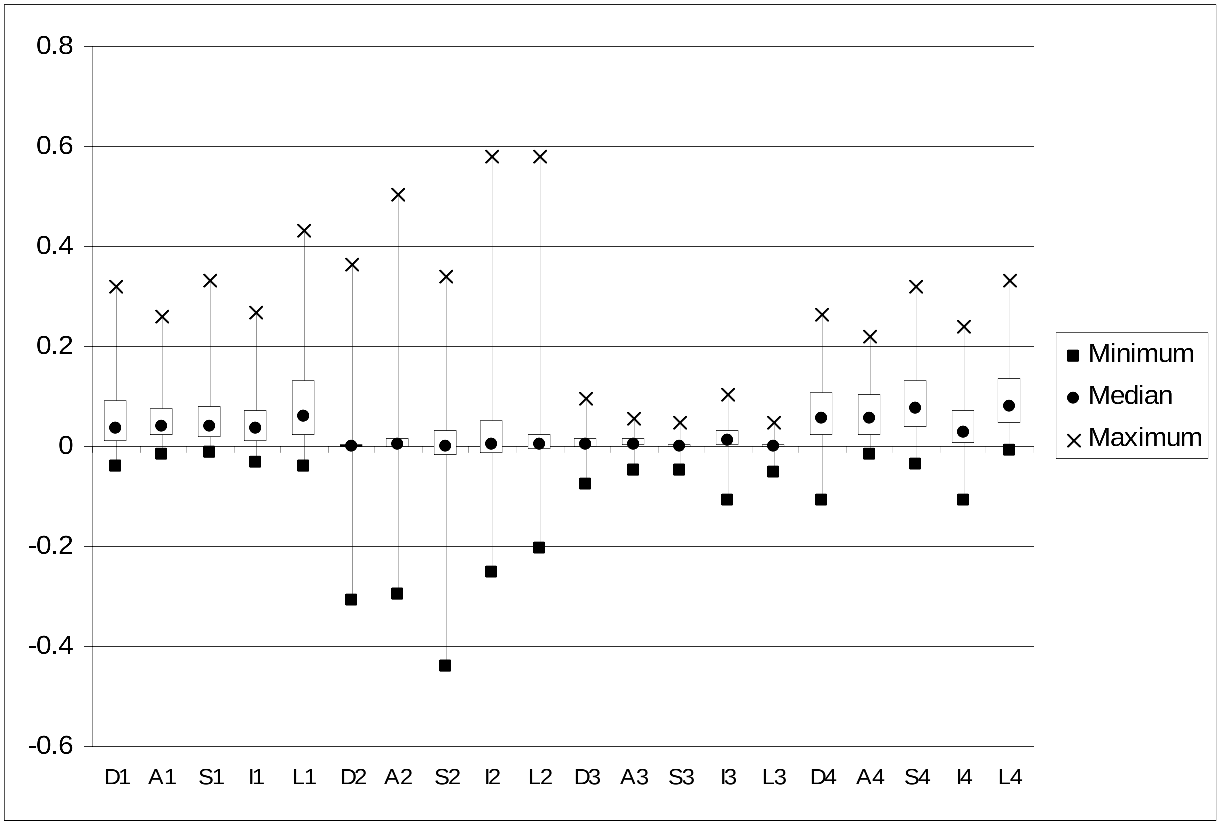

Figure 4: Box Charts for the Data Cleaning Situation

3.2 Discussion

The results in both experiments indicate the complex pattern of algorithm performance in the presence of asymmetric attribute noise. In general, over representative training noise should be avoided while under representative training noise is less of a concern. However, the interactions among algorithm, noise level, and training set size indicate that these general results may not apply to particular practice situations. Thus investments to obtain training data with representative noise levels may not be worthwhile in practice. Investments to clean field data seem more likely to be worthwhile consistent with the results in (Zhu and Wu 2004).

Two minor surprises involved the lack of impact of noise variation and ensemble learning algorithms. Although we think that variable attribute noise is the most realistic approach, it was surprising not to see that uniform and importance sampled noise had little differential impact. The results did not show much evidence that ensemble methods (AdaBoost-M1) had less relative sensitivity than other learning algorithms.

The results of these experiments are consistent with the results in (Perlich et al. 2003) about training set size and data set difficulty. Learning efficiency varies by classification algorithm. Similarly, the interaction of training set size and asymmetric attribute noise varies by algorithm. Data set difficulty as evidenced by average AUC scores with clean data seems to influence the effect of asymmetric attribute noise. However, we did not evaluate enough data sets to make any definitive conclusions about the impact of data set difficulty.

4. Conclusion

We presented an empirical comparison about asymmetric noise levels between training and field environments. We developed an innovative experimental design with algorithm, noise level, and training set size as factors and relative performance change as the performance measure. We considered under representative training noise (low training noise and high test noise) and over representative training noise (high training noise and low test noise). For noise generation, we studied uniform input noise levels on all attributes, variable noise levels, and noise levels assigned by attribute importance. Our results indicated that over representative training noise should be avoided while under representative training noise is less of a concern. However, the interactions among algorithm, noise level, and training set size indicate that these general results may not apply to particular practice situations.

This study with an emphasis on internal validity has limitations on conclusions in a wide variety of domains. To study the interactions of noise level differences and training set size, internal validity with controls about confounding effects was necessary. A follow-on experiment with an emphasis on data set characteristics such as difficulty and prevalence would provide additional external validity to complement this study. In addition, variations of learning algorithms could be studied to understand the impact of noise handling methods on asymmetric attribute noise. We are also interested in similar experiments to study class noise.

References

Aha, D., and Kibler, D., "Instance-based learning algorithms," *Machine Learning* 6, 1991, 37 – 66.

Cessie, S. and Houwelingen, J., "Ridge Estimators in Logistic Regression," *Applied Statistics* 41, 1 (1992), 191 – 201.

Dietterich, T., "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning* 40, 2 (2000), 139 – 158.

Freund, Y. and Schapire, R., "Experiments with a new boosting algorithm," in *Proc. International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, 1996, pp. 148 – 156.

Goldman, S. and Stone, R., "Can PAC Algorithms Tolerate Random Attribute Noise?," *Algorithimica*, 14 (1995), 70 – 84.

Hettich, S., Blake, C., and Merz, C., UCI Repository of Machine Learning Databases *(http://www.ics.uci.edu/~mlearn/MLRepository.html),* Department of Information and Computer Science, University of California, Irvine, 1998.

Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design," *Neural Computation* 13, 3 (2001), 637 – 649.

Laird, P. *Learning from Good and Bad Data*, Kluwer Academic Publishers, Norwell, MA, 1988.

Orr, K., "Data Quality and Systems Theory," *CACM* 41, 2 (February 1998), 66 – 71.

Perlich, C., Provost F., and Simonoff, J., "Tree Induction vs. Logistic Regression: A Learning-curve Analysis," *Journal of Machine Learning Research* 4 (2003), 211 – 255.

Pierce, D. and Ackerman, L., "Data Aggregators: A Study of Data Quality and Responsiveness," available from *http://www.privacyactivism.org/docs/DataAggregatorsStudy.html*, May 2005.

Quinlan, J. "Induction of Decision Trees," *Machine Learning*, 1 (1986), 81 – 106.

Quinlan, J. "The Effect of Noise on Concept Learning," in *Machine Learning, an Artificial Intelligence Approach*, Volume II, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), Morgan Kaufmann, 1986, 149 – 166.

Quinlan, R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Redman, T., *Data Quality for the Information Age*, Artech House, 1996.

Redman, T., "The Impact of Poor Data Quality on the Typical Enterprise," *CACM* 41, 2 (February 1998), 79 – 82.

Witten, I. and Frank, E., *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Zhu, X. and Wu, X., "Class noise vs. attribute noise: a quantitative study of their impacts," *Artificial Intelligence Review* 22, 3 (2004), 177 – 210.